Sirius Tutorial

Speakers: Jason Mars, Lingjia Tang, Johann Hauswald, Yiping Kang, Yunqi Zhang, Michael A. Laurenzano
Tutorial Schedule

- **Part 1: Introduction of Sirius**
  - *Jason Mars, Lingjia Tang*
- **Part 2: Sirius Demo**
  - *Johann Hauswald*
- **Break**
- **Part 3: Sirius-Suite**
  - *Johann Hauswald, Yiping Kang, Yunqi Zhang*
- **Part 4: DNN: an Emerging IPA Workload**
  - *Michael A. Laurenzano*
- **Closing Remarks, Questions and Discussion**
  - *Everyone*
Sirius Tutorial

Topic: Introduction
Speakers: Jason Mars, Lingjia Tang
Sirius: An Open End-to-End Voice and Vision Personal Assistant

Papers

2015


jasonmars.org

lingjia.org

clarity-lab.org
Sirius: An Open End-to-End Voice and Vision Personal Assistant
“Wearable Computing Devices, Like Apple iWatch, Will Exceed 485 Million Annual Shipments by 2018”
- ABI Research

Questions Arise

• What are the impacts on our data centers?
• Can we scale to millions (or billions) of users?
• How do we design the right systems?
Questions Arise

• What are the impacts on our data centers?
• Can we scale to millions (or billions) of users?
• How do we design the right systems?

"Wearable Computing Devices, Like Apple iWatch, Will Exceed 485 Million Annual Shipments by 2018" - ABI Research

The Hard Place: Build an end to end intelligent personal assistant.
Siri(us) - End-to-End Voice and Vision Personal Assistant
**Siri(us)** - End-to-End Voice and Vision Personal Assistant

- **Call my doctor.**
  - Voice Command (VC)

- **Who is the lead signer of U2?**
  - Voice Query (VQ)

- **When does this bar close?**
  - Voice/Image Query (VIQ)
The life of a query through the Sirius Pipeline
Call my doctor.
Call my doctor.
Call my doctor.
Call my doctor.
Call my doctor.
Call my doctor.
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Who is the lead signer of U2?
Who is the lead signer of U2?
Who is the lead signer of U2?
Who is the lead signer of U2?
Who is the lead signer of U2?
Who is the lead signer of U2?
Who is the lead signer of U2?
Sirius: An Open End-to-End Voice and Vision Personal Assistant
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
When does this bar close?
Sirius: An Open End-to-End Voice and Vision Personal Assistant

Users

Voice Command (VC)  Voice Query (VQ)  Voice-Image Query (VIQ)

Query Taxonomy

Sirius Hierarchy

IPA Services
Algorithmic Components
Automatic-Speech Recognition (ASR)
Stemmer
Regular Expression
Conditional Random Fields
Feature Extraction
Feature Description
Question Answering (QA)
Image Matching (IMM)
Sirius: An Open End-to-End Voice and Vision Personal Assistant
SIRIUS Hierarchy

**Query Taxonomy**
- Voice Command (VC)
- Voice Query (VQ)
- Voice-Image Query (VIQ)

**IPA Services**
- Automatic-Speech Recognition (ASR)
- Question Answering (QA)
- Image Matching (IMM)

**Algorithmic Components**
- HMM/GMM or HMM/DNN
- Regular Expression
- Stemmer
- Conditional Random Fields
- Feature Extraction
- Feature Description

SIRIUS: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant

**ASR**

- Speech Decoder
  - Feature Extraction
  - Feature Vectors
  - Speech Decoder
    - Input Layer
    - Hidden Layers
    - Output Layer
  - DNN Scoring or GMM scoring
  - Trained Data
    - Acoustic Model
    - Language Model
    - Word Dictionary
  - “Who was elected 44th President?”

**IMM**

- SURF Feature Extraction
  - Calculate Hessian Matrix
  - Build Scale-Space
  - Find Keypoints
- Image Descriptors
  - Haar Wavelet
  - Orientation Assignment
  - Keypoint Descriptor
- Image
- Keypoints
- Descriptor Database

**QA**

- Question-Answering
  - Document Selector
  - Scores
  - Websearch
  - “Barack Obama”
- Input Filter
  - Regex
  - 44th
  - “...elected 44th president”
  - Stemmer
  - “ed”
  - elect
  - CRF
  - elected 44th president

**Three Horsemen of Sirius**
Sirius Tutorial

Topic: Sirius Demo
Speaker: Johann Hauswald
Sirius Tutorial

Topic: Sirius-suite
Speakers: Johann Hauswald, Yiping Kang, Yunqi Zhang
What is Sirius-suite?

• Suite of 7 workloads extracted from the end-to-end Sirius intelligent personal assistant

• Suite includes:
  • Standalone workloads
  • Pretrained models (when applicable)
  • Inputs for all workloads

• Freely available at:
  sirius.clarity-lab.org
How does Sirius work?

Query Taxonomy

IPA Services

Users

Voice Command (VC)

Voice Query (VQ)

Voice-Image Query (VIQ)

Automatic-Speech Recognition (ASR)

Question Answering (QA)

Image Matching (IMM)

CMU Sphinx

Ephyra

OpenCV

KALDI
How does Sirius work?

Query Taxonomy

IPA Services

Tasks

Signal Processing

Natural Language Processing

Image Processing
How does Sirius work?

Tasks

Signal Processing  Natural Language Processing  Image Processing  Tasks
### How does Sirius work?

<table>
<thead>
<tr>
<th>Signal Processing</th>
<th>Natural Language Processing</th>
<th>Image Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)</td>
<td>Regular Expression</td>
<td>Feature Extraction</td>
</tr>
<tr>
<td></td>
<td>Stemmer</td>
<td>Feature Description</td>
</tr>
<tr>
<td></td>
<td>Conditional Random Fields</td>
<td></td>
</tr>
</tbody>
</table>

**Tasks**

**Algorithmic Components**
How does Sirius work?

<table>
<thead>
<tr>
<th>Signal Processing</th>
<th>Natural Language Processing</th>
<th>Image Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)</td>
<td>Stemmer</td>
<td>Feature Extraction</td>
</tr>
<tr>
<td></td>
<td>Regular Expression</td>
<td>Feature Description</td>
</tr>
<tr>
<td></td>
<td>Conditional Random Fields</td>
<td></td>
</tr>
</tbody>
</table>

85% or 78%

**Tasks**

- Algorithmic Components
  - Stemmer
  - Conditional Random Fields
  - Feature Extraction
  - Feature Description
How does Sirius work?

<table>
<thead>
<tr>
<th>Signal Processing</th>
<th>Natural Language Processing</th>
<th>Image Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)</td>
<td>Stemmer, Conditional Random Fields, Regular Expression</td>
<td>Feature Extraction, Feature Description</td>
</tr>
</tbody>
</table>

85% or 78% 85%
How does Sirius work?

<table>
<thead>
<tr>
<th>Signal Processing</th>
<th>Natural Language Processing</th>
<th>Image Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)</td>
<td>Stemmer Conditional Random Fields</td>
<td>Feature Extraction Feature Description</td>
</tr>
<tr>
<td><strong>85% or 78%</strong></td>
<td><strong>85%</strong></td>
<td><strong>97%</strong></td>
</tr>
</tbody>
</table>

Tasks

Algorithmic Components
How does Sirius work?

Tasks

- Signal Processing
- Natural Language Processing
- Image Processing

Algorithmic Components

- Gaussian Mixture Model (GMM)
- Deep Neural Network (DNN)
- Regular Expression
- Stemmer
- Conditional Random Fields
- Feature Extraction
- Feature Description

7 kernels: 92% of total execution of Sirius

85% or 78%
85%
97%
Objectives of Sirius-suite

- Represent emerging datacenter workloads
  - speech recognition, computer vision, natural language processing

- Cover key components in the application
  - Suite represents 92% of the total cycles

- Implementations
  - Single thread CPU
  - Pthread CPU
  - GPU

- Easy to use
  - Implemented in C/C++ and CUDA
## Sirius-suite

<table>
<thead>
<tr>
<th>Service</th>
<th>Workload</th>
<th>Task</th>
<th>Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Matching</td>
<td>Feature Extraction</td>
<td>Computer Vision</td>
<td>Pthread, GPU</td>
</tr>
<tr>
<td></td>
<td>Feature Description</td>
<td>Computer Vision</td>
<td>Pthread, GPU</td>
</tr>
<tr>
<td>Automatic Speech Recognition</td>
<td>GMM Scoring</td>
<td>Speech Recognition</td>
<td>Pthread, GPU</td>
</tr>
<tr>
<td></td>
<td>DNN Scoring</td>
<td>Speech Recognition</td>
<td>Pthread, GPU</td>
</tr>
<tr>
<td>Question Answer</td>
<td>Stemmer</td>
<td>Natural Language Processing</td>
<td>Pthread, GPU</td>
</tr>
<tr>
<td></td>
<td>Regular Expression</td>
<td>Natural Language Processing</td>
<td>Pthread</td>
</tr>
<tr>
<td></td>
<td>Conditional Random Fields</td>
<td>Natural Language Processing</td>
<td>Pthread</td>
</tr>
</tbody>
</table>
Image Matching

Speaker: Johann Hauswald
Sirius: An Open End-to-End Voice and Vision Personal Assistant

- **Image**
  - Calculate Hessian Matrix
  - Build Scale-Space
  - Find Keypoints

- **Keypoints**
  - SURF Feature Extractor
  - SURF Feature Descriptor
    - Haar Wavelet
    - Orientation Assignment
    - Keypoint Descriptor

- **Image Descriptors**
  - Descriptor Database

- **Diagram**
  - FE (41%)
  - ANN (3%)
  - FD (56%)
Accelerated workload: 97%
Speeded Up Robust Features (SURF)

- **Feature Extraction:**
  - Image down sampled and convolved multiple times to find interesting regions
  - Convolution responses are thresholded
  - Local maxima are kept as keypoints

- **Feature Description:**
  - Keypoints are assigned an orientation vector and grouped with other similar keypoints
  - Each group of keypoints forms a descriptor
  - Robust to scale, rotation, illumination
Feature Extraction (FE) Overview

• Extracts keypoints from image using OpenCV’s SURF
• Input set includes images of various sizes
• Multithreaded parallelism assigns image tiles to each thread
• GPU implementation is highly optimized OpenCV
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Sirius: An Open End-to-End Voice and Vision Personal Assistant
Feature Description (FD) Overview

- Generates image descriptors using previously extracted keypoints
- Descriptors offer a robust representation of the image to match against other images.
- Multithreaded version computes descriptors from image tiles
- GPU implementation is highly optimized OpenCV
Automatic Speech Recognition

Speaker: Yiping Kang
Automatic Speech Recognition (ASR)

Pre-processing

Speech

Feature Extraction

Feature Vectors

Input Layer

Hidden Layers

Output Layer

DNN

GMM

HMM

Viterbi Algorithm

cap, i, t, al
Automatic Speech Recognition (ASR)

Recognition Process

Input Layer  Hidden Layers  Output Layer

Sirius: An Open End-to-End Voice and Vision Personal Assistant
Automatic Speech Recognition (ASR)

Recognition Process

- DNN: 78% (HMM: 22%)
- GMM: 85% (HMM: 15%)
Gaussian Mixture Model (GMM)

- GMM maps an input feature vector into a multi-dimensional space and model the posterior distribution of HMM states
- Speech feature vector are input set
- Produce posterior probabilities for HMM states
- Multithreaded implementation scores features in parallel
- GPU parallelism similar to multithreaded implementation on CPU
Deep Neural Network (DNN)

- “Deep”
  - An artificial neural network with multiple hidden layers
Deep Neural Network (DNN)

Input Layer

Inner Product Layer
\[ \text{out} = \text{in} \times \text{weight} + \text{bias} \]

Sigmoid Layer
\[ \text{out} = \text{sigmoid}(\text{in}) \]

Sigmoid Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Inner Product Layer

Feature Vector

7 Fully Connected Layers
6 Activation Layers

\[ f(x) = \frac{1}{1 + e^{-x}} \]

Probabilities

7 Fully Connected Layers
6 Activation Layers

\[ f(x) = \frac{1}{1 + e^{-x}} \]

\[ \text{sigmoid} \]

\[ \text{Probabilities} \]
Deep Neural Network (DNN)

- Input set includes speech feature vectors
- Produce posterior probabilities for HMM states
- Multithreaded parallelism perform matrix operation in parallel
- GPU implementation is state-of-the-art open source neural network library (Caffe)
Question Answering

Speaker: Yunqi Zhang
Question Answering

Who was elected 44th president of the United States?

Input Filters
Candidate Phrases:
• (*) was elected …
• (*) was the 44th …

Document Database
e.g., wikipedia

Document Filters

Answer Selector
Barack Obama

Sentences in Candidate Documents:
• wikipedia/Barack_Obama
• wikipedia/Presidents_of_the_US
Question Answering

85% of the execution
- Conditional Random Fields
- Stemmer
- Regular Expression

46% 12% 17% 22% 3%
- Stemmer
- Other
- CRF
- Search
- Regex

Candidate Phrases:
• (*) was elected ...
• (*) was the 44th ...

Candidate Documents:
- wikipedia/Barack_Obama
- wikipedia/Presidents_of_the_US
Stemmer

Reduce inflected (or derived) words to their word stem, base or root form

Who was elected 44th president of the United States?

be elect presid state
Stemmer

Who was elected 44th president of the United States?
Who was elected 44th president of the United States?
Stemmer

- who
- was
- ... ...
- states
Regular Expression

Who was elected 44th president of the United States?

e.g., extracting what the question is asking for

• (what|which|name|give|tell) (.*)?capital
  capital city

• (what|which|name|give|tell) (.*)?speed
  speed

• how (close|deep|far|high|long|narrow|short|tall|wide)
  dimension

• ......

• who(m|se)
  person
Regular Expression

Who was elected 44th president of the United States?
Who was elected 44th president of the United States?

- (what|which|name|give|tell) (.* )?capital
- how (close|deep|far|high|long|narrow|short|tall|wide)
- . . . .
- who(m|se)?
Regular Expression

(what|which|name|give|tell) (.*)?capital

how (close|deep|far|high|long|narrow|short|tall|wide)

......

who(m|se)?
Conditional Random Fields

Part of Speech Tagging

- pronoun.
- verb.
- noun.
- adverb.

Who was elected 44th president of the United States?

- verb.
- num.
- adverb.
- noun.

- common / proper noun.
- singular / plural
- whether possessive
- .......
Conditional Random Fields

Who was elected 44th president of the United States?
Who was elected 44th president of the United States?

- **wiki/Barack_Obama**
  Barack Hussein Obama II is the 44th and current President of the United States

- **wiki/List_of_Presidents_of_the_United_States**
  … and 44th president, Barack Obama, is the first president of African descent

- **wiki/United_States**
  Barack Obama, the first African American, and 44th president, was elected in 2008 …
Conditional Random Fields

wiki/Barack_Obama
Barack Hussein Obama II is the 44th and current President of the United States

wiki/List_of_Presidents_of_the_United_States
… and 44th president, Barack Obama, is the first president of African descent

wiki/United_States
Barack Obama, the first African American, and 44th president, was elected in 2008 …
Running Sirius-suite
Run-to-Run Variation

- Low variance
- High reproducibility

IMM

Feature Extraction

GMM Automatic Speech Recognition

ASR

Feature Description

DNN Automatic Speech Recognition

QA

Regular Expression

Stemmer

Conditional Random Fields
Speedup

**IMM**

**ASR**

**QA**

**CPU specification**
- Intel Xeon E5-2630v3 @ 2.40GHz
- 1 socket * 8 cores w/o SMT

**GPU specification**
- Nvidia K40
- 2880 processor cores
Deep Neural Networks: an Emerging IPA Workload

Speaker: Michael Laurenzano
Why DNN?

• Starting to emerge in a number of domains
  • Big data analytics
  • Video/image/audio recognition
  • Language semantics/translation
• Many components of Sirius could be done with DNN
• Why not DNN?
  • Intense computation and data movement
  • The dark art of configuration
  • Need lots of training data
10,000 foot view of DNN

- What is “deep” about a DNN?
  - More layers
  - Often, bigger layers also
- Example — DNN in Sirius ASR

\[
p(HMM\text{state} | \text{AcousticModel})
\]
Sirius and DNN

- Several parts of the pipeline could use DNN
Studying DNNs as an architect is a challenge

- What should my DNN do?
- Network configuration
  - How many layers?
  - How to configure those layers?
    - Interconnections?
    - Activation functions?
- Model training
  - When are the results accurate enough?
  - How to choose a training set?

Need a common infrastructure to study DNN behavior
Introducing Djinn

- Djinn
  - Infrastructure for DNN as a service
  - Extensible to a number of DNN services
  - CPU-optimized implementation
  - GPU-optimized implementation

- Djinn-suite
  - 7 complete DNN services covering important IPA use cases
  - Configured networks and trained models are provided
    - No machine learning expertise required
Djinn-suite

- Image Processing
  - Image classification (IMC)
  - Facial recognition (FACE)
  - Digit recognition (DIG)
- Speech Processing
  - Automatic speech recognition (ASR)
- Natural Language Processing
  - Part-of-speech tagging (POS)
  - Chunking (CHK)
  - Name entity recognition (NER)
Basic characterization

• Single CPU

Not all DNN services are created equal

Accelerating DNN can provide significant speedup
DNN on the GPU

- Out of the box Caffe (cuDNN — optimized CUDA version of Caffe)
- Single GPU
Now let’s optimize for throughput

- Throughput improvement techniques
  - Batch queries to the GPU — fewer transfers
  - Multiple DNN instances via Multi Process Service (MPS)
Remaining bottlenecks?

- Multiple GPUs (up to 8)
Stay tuned…

- Djinn paper will appear at ISCA 2015

- Djinn is for you!
  - Software released with paper
  - Configurable Djinn service
  - 7 pre-packaged Djinn-suite applications
Sirius Tutorial

Topic: Closing Remarks
Speaker: Johann Hauswald